Proc. Int. Workshop on Human Memory-Inspired Multimedia Organization and Preservation (HMMP) at the IEEE Int. Conf. on Multimedia and Expo (ICME 2015), June-July 2015, Torino, Italy

PHOTO COLLECTION CONTEXTUALIZATION

Konstantinos Apostolidis, Vassilios Solachidis, Olga Papadopoulou, Vasileios Mezaris

Information Technologies Institute, CERTH, Thessaloniki, Greece
{kapost, vsol, olgapap, bmezaris}@iti.gr

ABSTRACT

We propose a photo collection contextualization method that enriches the content of a user collection with additional photos extracted from collections of different users that attended the same event. Depending on the user needs, the selected photos may depict scenes from either the same or different aspects of the event. To achieve contextualization, we combine techniques for photo clustering to sub-events, multi-user time synchronization, and sub-event matching, utilizing visual, time and geolocation information. The proposed method has been tested on datasets provided by MediaEval 2014 SEM Task.

Index Terms— photo similarity, sub-event, contextualization, personal photo collection

1. INTRODUCTION

People are used to capture photos when attending events (e.g. trips, concerts, celebrations, etc.), store and retrieve them after a short or long period of time in order to refresh their memories, present their experience to others or just look back to a nice remembrance.

In the past, when the use of capturing devices was limited, people remembered and described an event just by their memories. However human memory tends to delete details. This is more intense for episodic memory [1] which includes specific events that took place at a particular place and time. Forgetting from episodic memory is rapid and substantial. On the other hand, forgetting from semantic memory, which is responsible for knowledge and skills, is much less rapid and information is well preserved over long periods or never lost.

With the widespread use of digital media capture devices (mobile phones, digital cameras, tablets), people tend to acquire a multitude of photos in short time resulting in large collections for each attended event consisting of photos, significant or not. Due to this, several works have been introduced in the literature aiming to organize [2], visualize [3] or summarize [4] photo collections in order to distinguish the valuable information for the user, especially for future reference. On top of these, we propose a method that enriches the content of a photo collection.

Contextualization is defined as the process of providing information about a situation in which something happens. Contextualization plays a critical role in how and when individuals remember the past, since contextualization information may help form well preserved semantic memories of an event.

Events that a person attends can be either *public* events (e.g. sports competitions, concerts, etc.) or *personal* events (e.g. a professional meeting, a birthday party, a family trip, etc.). Contextualization can be applied in both cases. To give a more specific example, suppose that a person attends a *public* event such as some competitions of the Olympic games of London 2012 and takes a number of photos. At the end of the event, what remains to the user are the memories and the photos. Huge number of people attended this event capturing the sports and more specifically the moments which they considered interesting. The information collected by individuals is diverse covering the overall event. Combining this information and sharing it, we manage to provide a more holistic description of the event which can later trigger the memory of its attendant. On the other side, an example of a *personal* event can be a project meeting. All meeting participants own a device to capture their photos and at the end of the meeting each of them has created a photo collection depicting their moments of interest. By applying the contextualization method to a photo collection of an individual, we improve the recall of the event using the photos of his colleagues.

We propose a method that, given a photo collection, it enriches its content by retrieving additional content (photos covering different aspects of the same event) from other resources resulting in a more complete view of the event. The rest of the paper is organized as follows: in Section 2 related work is presented, in Section 3 our method is described while in Sections 4 and 5 the experimental results and conclusions are given.

2. RELATED WORK

To the best of our knowledge, this work is the first that deals with personal photo collection contextualization. However, a variety of methods have been proposed in the literature which tackle the problem of organizing and visualizing personal photo collections [5, 6].

Visual information is employed to assess the similarity of photos in a collection. Both global and local visual descriptors are used for comparing the visual content as presented in [7], as well as high level representation of photos for comparing photos semantically [4].

In [8], a method which compares events at collection level instead of taking into account individual photos features is described. Given an event described by a set of photos, this method retrieves photos from the same event captured by different users, utilizing visual, temporal as well as geolocation information, although the authors do not test geolocation information in the conducted experiments.

Existing methods which select one representative photo

This work was supported by the European Commission under contract FP7-600826 ForgetIT.

out of a photo collection deal mainly with summarization problems. Summarizing a photo collection [4], a small subset of the total photos is selected as representatives, offering an overview of the entire collection.

3. PROPOSED METHOD

The proposed method consists of two steps, the pre-processing step and the actual contextualization step as illustrated in Fig. 1.

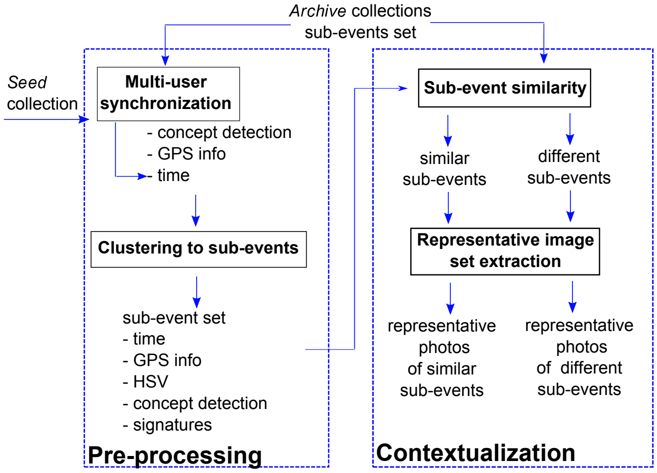


Fig. 1. Method overview

Let *seed* be the user collection to be contextualized and *archive* the ones of the other users referring to the same event. We assume that seed and archive collections belong to the same event, thus, we are not dealing with recognizing the event. All collections consist of *sub-events*, namely groups of photos that constitute a distinct action within a bounded time and space. In the Olympic games event example, the individual sports competitions (e.g. first day of tennis men, weightlifting women final, etc.) can be considered as sub-event of the entire London 2012 summer Olympic event.

In the pre-processing step, we cluster all the collections into sub-events. The clustering to sub-events method employs: time information (photo capture date and time), geolocation (if available) and concept detection scores (extracted by applying a concept detection algorithm [9]). Using multiple features is particularly important in the case of long events with multi-location or temporally overlapping sub-events.

Although time can be easily extracted from photo EXIF tag, further processing may be required since the photos have been captured from different devices and photo collections may not be temporally synchronized. Therefore, before applying the clustering to sub-events, we use the multi-user time synchronization method presented in [10].

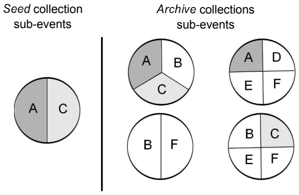


Fig. 2. Clustering to sub-events example of seed and archive collections

An example of clustering result is depicted in Fig. 2. The proposed method enriches the content of the seed collection with a user-defined number of photos from archive collections. It should be noted that representative photos from both similar (A and C sub-events) and different sub-events (B, D, E and F) are selected. The percentage of photos taken from similar and different sub-events is user-defined.

In the rest of this section we present in detail solely the actual contextualization step, which consists of the sub-event matching and the seed collection enrichment.

3.1. Sub-event matching

We establish distance measures that are employed to match a seed collection sub-event to sub-events of the archive collections. Let S be the seed collection that is clustered in N_S sub-events $S^i, i = 1, ..., N_S$. Each sub-cluster S^i contains N_{S^i} photos $s^{i,j}, j = 1, ..., N_{S^i}$. For each photo $s^{i,j}$ we can use the HSV histogram $s^{i,j}_{HSV}$, the GPS location $s^{i,j}_{GPS}$, the concept detection scores $s^{i,j}_{SC}$ and the capture time $s^{i,j}_{T}$ information. Similarly, $A^j_k, j = 1, ..., N_{A_k}$ is the $j-th$ sub-event of the archive collection of user k denoting the corresponding photo information as $a^{i,j}_{k,\{HSV/GPS/SC/T\}}$. We introduce four sub-event matching approaches.

time: The baseline approach utilizes the time information of the photos. The distance between a seed sub-event and an archive one is the minimum pairwise temporal distance of their photos and it is given by:

$$D_T(S^i, A^j_k) = \begin{cases} \min\limits_{m,n}\{|(s^{i,m}_T - a^{j,n}_{k,T}|)\}, & \text{if sub-events are not temporally overlapping} \\ 0, & \text{otherwise} \end{cases}$$

where $1 \leq m \leq N_{S^i}$ and $1 \leq n \leq N_{A^j_k}$.

HSV: In this approach, sub-event distance is defined as the minimum pairwise distance between the HSV color histograms of the sub-events photos:

$$D_H(S^i, A^j_k) = \min_{m,n}\{d_{\cos}(s^{i,m}_{HSV}, a^{j,n}_{k,HSV})\}$$

where $d_{\cos}$ is the cosine distance.

scores: In this case, we use the minimum pairwise distance between vectors of concept detection scores:

$$D_{SC}(S^i, A^j_k) = \min_{m,n}\{d_{cd}(s^{i,m}_{SC}, a^{j,n}_{k,SC})\}$$

where d_{cd} is a distance used for concept detection scores, given by Eq. (6) in [11].

signature: While the aforementioned approaches employ pairwise photo based distances, we adopt an approach that calculates the distances among entire sub-events. Each sub-event is described by a single vector called signature. To construct the signature vector of a sub-event we calculate the mean and variance of vectors of concept detection scores. We then keep the u concept indices with the highest mean and finally we sort them in ascending order of their variances.

The signature vector is representative of the entire sub-event since it contains concepts that attained high score (top u) while at the same time, their position in the vector shows the significance of the specific concept (low variance).

The indices of the concept c of sub-events S^i and A^j_k are denoted as $S^i_{SIG_c}$ and A^j_{k,SIG_c} respectively. If the concept c does not belong to the top u concept indices, then its value is set to 0.

The distance of two sub-event signatures equals the sum of the distances D_C of the signatures for all concepts.

$$D_{SIG}(S^i, A^j_k) = \sum_{c=1}^{N_C} D_C(S^i_{SIG_c}, A^j_{k,SIG_c})$$

If a concept c belongs in both signatures, then distance D_C equals the $\mathrm{W}(p) \cdot p$ where $p = |(S^i_{SIG_c} - A^i_{k,SIG_c})|$ (1). It should be noted that W is a function used for weighting. As a result, concept indices with low variance have more impact on the distance calculation. The intuition of this distance measure is that two sub-events are close if they contain more common concepts having high score and same positions. If a concept index exists only in one signature, then D_C is set to $W(i)(u+1)$ (2), where i is the index of concept c in this signature. Finally, if a concept index does not exist in any of the signatures then D_C is set to $\mathrm{W}(p+1) \cdot (u+1)$ (3).

$$D_C(S^i, A^j_k) =$$

$$\begin{cases} \mathrm{W}(p)\cdot p, & \text{if } S^i_{SIG_c} > 0 \text{ and } A^j_{k,SIG_c} > 0 \quad (1) \\ \mathrm{W}(p)\cdot(u+1), & \text{if } S^i_{SIG_c} > 0 \text{ xor } A^j_{k,SIG_c} > 0 \quad (2) \\ \mathrm{W}(p+1)\cdot(u+1), & \text{otherwise} \quad (3) \end{cases}$$

There are cases where time information is not available (e.g. photos are captured using an old camera and no metadata are saved) or not exploitable (e.g. when photo capture devices of different users are not synchronized and multi-user time synchronization has not been applied). In these cases, only the last three sub-event distance measures (D_T, D_H, D_{SC}) can be applied. On the other hand, if time information is available and photo collections are synchronized, time-based sub-event distance measure can be combined with other sub-event distance measures. We also choose to test the weighted sum of time and HSV, scores or signatures $(D_{T,x} = w \cdot D_T + (1-w) \cdot D_{\{x\}})$ where $x \in \{HSV, SC, SIG\}$.

Geolocation information is employed to further refine the matched sub-events. We calculate a distance threshold, t_{GEO}, that denotes if two sub-events are considered spatially close or not. For the t_{GEO} estimation we cluster all pairwise archive photo spatial distances using the k-means clustering algorithm for number of clusters $k = 2$. Threshold t_{GEO} equals with the lowest cluster center. The pairs that are not spatially close are excluded from the matching procedure.

Finally, we select a subset of archive sub-events that their distance to a seed sub-event is below the corresponding thresholds $(t_T, t_{T,x}, t_x)$, where $x \in \{HSV, SC, SIG\}$.

3.2. Seed collection enrichment

Having matched each seed user sub-event to an archive sub-event, we select the photos of the matched archive sub-events that are either similar or dissimilar to the seed user collection photos. To achieve this, the proposed method uses two user-controlled parameters, a and b. Parameter a $(a > 0)$, controls the total number of photos that will be used for contextualizing the seed collection. Parameter b $(0 \leq b \leq 1)$, specifies what percentage of these photos should belong to sub-events that are also contained in the seed collection; the rest of the photos that will be added, will be chosen so as to belong to different sub-events.

We select the $a \cdot b \cdot N$ archive sub-event photos (where $N = \sum_{i=1}^{N_S} N_{S^i}$ is the total number of seed collection photos) that are most similar to the seed collection sub-events from each archive sub-event we pick one photo which is the most dissimilar to the photos already contained in this sub-event of the seed collection. The set of the $a \cdot b \cdot N$ photos is the set of photos that contextualizes the sub-events of the event that were originally contained in the seed collection. The remaining $a \cdot (1-b) \cdot N$ photos that enrich the seed collection are collected from the most dissimilar sub-events.

3.3. Significant sub-events detection

In the case of $b \approx 0$, the user requests the contextualization of his collection to be applied only with different sub-events and the selection is based on a measure of photo significance.

In [12], a method that detects significant events in personal photo collections is presented. More specifically, user's picture-taking behavior is modeled using time-series and points with high residuals are flagged as potential significant events. In the proposed method we follow a simpler approach in which we calculate the significance of each photo according to the temporal distance from the rest of the photos of the sub-event. The significance SG of a photo $a^{j,l}_k$ which belongs in sub-event A^j_k, is given by:

$$SG(a^{j,l}_k) = \sum_{p=1}^{N_{A^j_k}} \exp\left(-g|a^{j,l}_{k,T} - a^{j,p}_{k,T}|\right)$$

Tuning g parameter, the exponential can become significant only for photos that are temporally close while for the rest the exponential is negligible. As a result, the sum depends also on the number of temporally close photos to $a^{j,l}_k$. Thus, given a sub-event, the significance of all photos is initially calculated and then the one having the highest significance is selected. If the archive sub-event is dissimilar to the seed collection sub-events then this photo is used as the representative one for the seed collection enrichment. If it belongs to one of the seed collection sub-events and the selected photo is temporally far from the seed collection photos and is selected as the representative one for the seed collection, otherwise the photo with the second highest significance is examined. The procedure is continued until a temporally distant photo is detected.

4. EXPERIMENTS AND RESULTS

We used the two datasets of the MediaEval 2014 SEM task [13] for our method evaluation. The *Vancouver* dataset consists of 1351 photos capturing various sub-events of the Vancouver 2010 Winter Olympic Games which is split into 35 user collections. The *London* consists of 2124 photos capturing various sub-events of the London 2012 Olympic Games which is split into 37 user collections. We used the *Vancouver* dataset to estimate the threshold values ($t_T, t_{T,x}, t_x$, where $x \in \{HSV, SC, SIG\}$) for each sub-event distance calculation approach and used their estimated values on the *London* dataset, which are $t_T = 0.10$, $t_{T,HSV} = 0.27$, $t_{T,SC} = 1.05$, $t_{T,SIG} = 0.69$, $t_{HSV} = 0.05$, $t_{SC} = 0.8$, $t_{SIG} = 0.27$. In the experiments conducted on the *London* dataset, the first user's ($user1$) collection is considered as the seed collection while the remaining 36 users form the archive collections which are pre-processed. For the preprocessing method of clustering to sub-events we used the ground truth of MediaEval 2014 SEM task [13] in order to test the actual contextualization part of our method. We also performed a set of experiments considering a subset of user collections, which contain at least two sub-events, as seed collections and averaged the results. It is worth noting that geolocation information is available for a subset of photos.

Since, sub-event matching is a classification task, we evaluate the results of our approaches using the established precision (P), recall (R) and F-measure (F1) measures. Table 1 shows the results for each matching approach, using the first user of the London dataset as seed and the rest of the users as the archive collections, since the first user has the most extended collection of sub-events. Table 2 shows the averaged results, using a subset of users as seed.

Table 1. Different sub-event matching methods evaluation using $user1$ collection as the seed.

	P	R	F1
time	0.466	0.872	0.607
time+HSV	**0.894**	0.764	**0.824**
time+scores	0.765	0.750	0.757
time+signatures	0.677	**0.880**	0.765
HSV	0.840	0.568	0.677
scores	0.659	0.763	0.707
signatures	0.773	0.425	0.548

Table 2. Different sub-event matching methods evaluation. Results are averaged from using all users collections as the seed.

	P	R	F1
time	0.399	**0.872**	0.536
time+HSV	**0.702**	0.724	**0.702**
time+scores	0.685	0.681	0.666
time+signatures	0.587	0.675	0.611
HSV	0.567	0.447	0.462
scores	0.624	0.549	0.525
signatures	0.544	0.323	0.392

We observe in Table 1 that the baseline method does not perform well due to the London dataset containing temporally overlapping events. Combining time information with the other features a significant boost is achieved (compare *time+HSV*, *time+scores* and *time+signatures* to *HSV*, *scores* and *signatures* approaches respectively). Specifically, *time+HSV* approach achieved the highest F1 measure. Furthermore, in the case where the time information is not available, *scores* approach can distinguish sub-events better than the other proposed approaches. The aforementioned observations using the first user as seed are confirmed for the averaged case (Table 2).

We numerically evaluate the impact of using these photos for contextualization by examining the sub-events that they contain. Specifically, we calculate three evaluation measures:

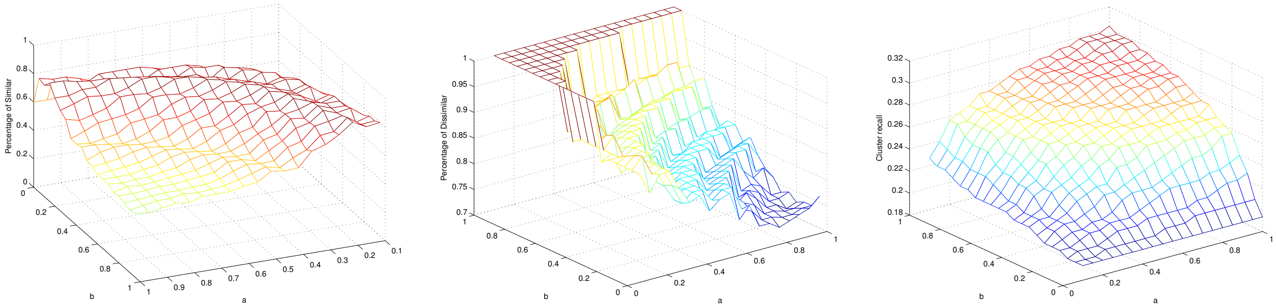
- **Percentage of similar (PoS)**: Out of the photos of the archive collections that were selected on the basis of representing similar sub-events, we measure the percentage of them that truly belong to such sub-events. This measure ranges from 0 to 1, where 1 is the optimal.
- **Percentage of dissimilar (PoD)**: Out of the photos of the archive collections that were selected on the basis of representing dissimilar sub-events, we measure the percentage of them that truly belong to such sub-events. This measure ranges from 0 to 1, where 1 is the optimal.
- **Cluster recall (CR)**: we measure the coverage increase after contextualization. Namely, the initial coverage of the seed collection is calculated based on the number of different sub-events of the total number of sub-events contained in the overall event. By contextualizing the collection we attempt to include more sub-events into the seed collection and increase the coverage.

Table 3. Percentage of the three evaluation measures for different a and b parameters

	a=0.5			a=1		
	b=0.2	b=0.5	b=0.8	b=0.2	b=0.5	b=0.8
PoS	0.8	0.76	0.57	0.83	0.58	0.46
PoD	0.82	0.84	0.9	0.72	0.81	0.88
CR	0.22	0.26	0.27	0.25	0.28	0.29

Fig. 3 illustrates the values of these measures when varying the values of parameters a and b, while indicative results for selected values of a and b are shown in Table 3. As far as the CR measure is concerned, the $user1$ seed collection consists of 46 sub-events out of the 238 total sub-events before contextualization, which is equal to 0.1932. As parameter a increases, meaning that the user selected to increase the number of photos that are used for contextualizing the seed collection, CR measure also increases and reaches almost 0.3. This indicates that the contextualized seed collection offers a broader coverage of the event, in comparison to the information contained in the seed collection prior to contextualization.

An example of the contextualization results, for both similar and dissimilar sub-events, which were selected by applying the proposed method on the *London* dataset are shown in Fig. 4 for visual inspection of the result. This example shows how the seed collection is contextualized with different photos. In Fig. 4a, the seed collection photos are illustrated, grouped in sub-events using ground truth. It seems that this collection contains photos from a part of the opening and award ceremonies, and the rowing (cox-less pair, eight, single scull, quad scull), weightlifting, soccer, marathon, long jump, races, wrestling, tennis, beach volley and judo competitions. Fig. 4b shows the photos of similar sub-events that were chosen from the archive collections for contextualization.



(a) (b) (c)

Fig. 3. Evaluation results using the three defined evaluation measures, for different a and b parameters (a) Percentage of Similar, (b) Percentage of Dissimilar and (c) Cluster recall

Finally, Fig. 4c shows the photos that were selected from the sub-events not present in the seed collection. These include photos from the taekwondo, cycling, fencing, basketball and horse riding competitions, as well as different parts of the opening and award ceremonies.

5. CONCLUSION

In this paper, we proposed a contextualization method that enriches the content of a user photo collection using photos from different user collections that attended the same event. The experimental results demonstrate the importance of using high level features extracted from photos or combining low level features with time information (if available) in contextualizing the user photo collection taking into account the way people preserve their memories of a specific event.

6. REFERENCES

[1] A. Baddeley, *Memory.* East Sussex New York, NY: Psychology Press, 2015.

[2] A. Ceroni, V. Solachidis, C. Niederée, O. Papadopoulou, N. Kanhabua, and V. Mezaris, "To keep or not to keep: An expectation–oriented photo selection method for personal photo collections," in *Proc. ACM Int. Conf. on Multimedia Retrieval (ICMRÂŠ15), Shanghai, China*, 2015.

[3] G. Tómasson, H. Sigurþórsson, B. Þ. Jónsson, and L. Amsaleg, "Photocube: Effective and efficient multi-dimensional browsing of personal photo collections," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval.* ACM, 2011, p. 70.

[4] C. Papagiannopoulou and V. Mezaris, "Concept-based image clustering and summarization of event-related image collections," in *Proceedings of the 1st ACM International Workshop on Human Centered Event Understanding from Multimedia.* ACM, 2014, pp. 23–28.

[5] A. Loui and A. Savakis, "Automated event clustering and quality screening of consumer pictures for digital albuming," *Multimedia, IEEE Transactions on*, vol. 5, no. 3, pp. 390–402, Sept 2003.

[6] J. Chen and S. Hibino, "Reminiscing view: Event-based browsing of consumer's photo and video-clip collections," in *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*, Dec 2008, pp. 23–30.

[7] L.-W. Kang, C.-Y. Hsu, H.-W. Chen, C.-S. Lu, C.-Y. Lin, and S.-C. Pei, "Feature-based sparse representation for image similarity assessment," *Multimedia, IEEE Transactions on*, vol. 13, no. 5, pp. 1019–1030, 2011.

[8] M. R. Trad, A. Joly, and N. Boujemaa, "Large scale visual-based event matching," in *Proceedings of the 1st ACM International Conference on Multimedia Retrieval*, ser. ICMR '11. New York, NY, USA: ACM, 2011, pp. 53:1–53:7. [Online]. Available: http://doi.acm.org/10.1145/1991996.1992049

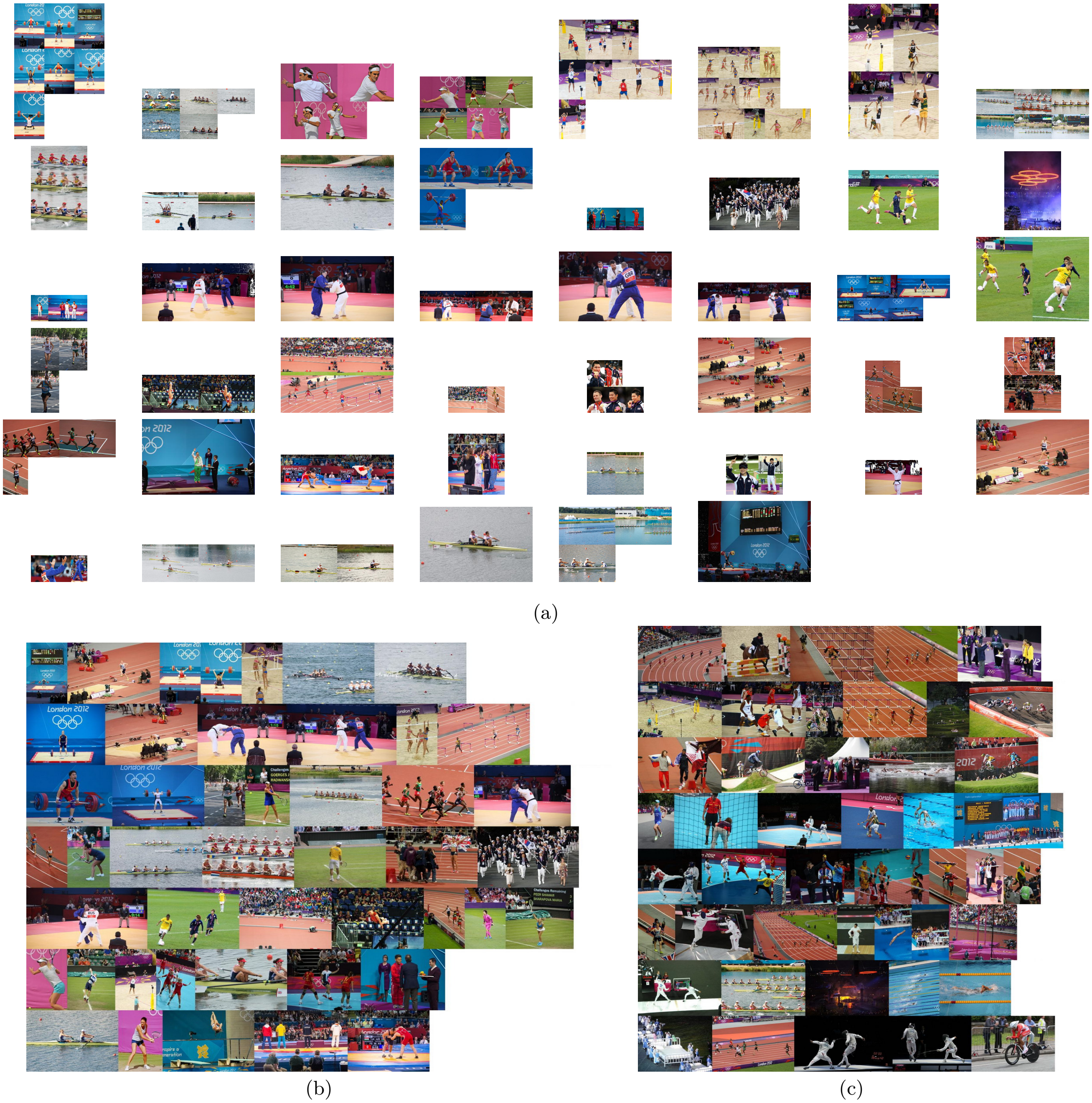
[9] F. Markatopoulou, N. Pittaras, O. Papadopoulou, V. Mezaris, and I. Patras, "A study on the use of a binary local descriptor and color extensions of local descriptors for video concept detection," in *MultiMedia Modeling*, ser. Lecture Notes in Computer Science. Springer International Publishing, 2015, vol. 8935, pp. 282–293. [Online]. Available: http://dx.doi.org/10.1007/978-3-319-14445-0_25

[10] K. Apostolidis, C. Papagiannopoulou, and V. Mezaris, "CERTH at MediaEval 2014 synchronization of multi-user event media task," in *Proceedings of the MediaEval Workshop*, 2014.

[11] V. Mezaris, P. Sidiropoulos, A. Dimou, and I. Kompatsiaris, "On the use of visual soft semantics for video temporal decomposition to scenes," in *Semantic Computing (ICSC), 2010 IEEE Fourth International Conference on*, Sept 2010, pp. 141–148.

[12] M. Das and A. Loui, "Detecting significant events in personal image collections," in *Semantic Computing, 2009. ICSC '09. IEEE International Conference on*, Sept 2009, pp. 116–123.

[13] N. Conci, F. D. Natale, and V. Mezaris, "Synchronization of multi-user event media (sem) at mediaeval 2014: Task description, datasets, and evaluation," in *MediaEval 2014 Workshop, Barcelona, Spain*, 2014.



(a)

(b)

(c)

Fig. 4. (a) Seed collection sub-events, (b) Photos added through contextualization that belong to sub-events already contained in the seed collection, (c) Photos added through contextualization that belong to different sub-events of the same event